Minimax Estimation of the Volume of a Set under the Rolling Ball Condition

Ery Arias-Castro, Beatriz Pateiro-López, Alberto Rodríguez-Casal

Version: Accepted Manuscript
This is an Accepted Manuscript of an article published by Taylor & Francis in Journal of the American Statistical Association on 2018, available online: http://www.tandfonline.com/10.1080/01621459.2018.1482751

HOW TO CITE

Ery Arias-Castro, Beatriz Pateiro-López & Alberto Rodríguez-Casal (2018) Minimax Estimation of the Volume of a Set Under the Rolling Ball Condition, Journal of the American Statistical Association, DOI: 10.1080/01621459.2018.1482751.

FUNDING

Research has been funded by project INNPAR2D (reference code MTM2016-76969-P), from the Ministry of Economy and Competitiveness and the European Regional Development Fund (ERDF).

Minimax Estimation of the Volume of a Set under the Rolling Ball Condition *

Ery Arias-Castro
Department of Mathematics, University of California
and
Beatriz Pateiro-López
Departamento de Estatística, Análise Matemática e Optimización,
Universidade de Santiago de Compostela
and
Alberto Rodríguez-Casal
Departamento de Estatística, Análise Matemática e Optimización,
Universidade de Santiago de Compostela

March 20, 2018

Abstract

We consider the problem of estimating the volume of a compact domain in a Euclidean space based on a uniform sample from the domain. We assume that the domain has a boundary with positive reach. We propose a data splitting approach to correct the bias of the plug-in estimator based on the sample α-convex hull. We show that this simple estimator achieves a minimax lower bound that we derive. Some numerical experiments corroborate our theoretical findings.

Keywords: minimax lower bound; r-convex hull; rolling condition; support estimation; volume estimation.

*This work was partially supported by the US National Science Foundation (DMS 1513465) and by the Spanish Ministry of Economy and Competitiveness and ERDF funds (MTM2016-76969P)

1 Introduction

We consider the problem of estimating the volume of a compact domain[1] S of a Euclidean space based on an IID sample from the uniform distribution supported on S. Concretely, we are given a set of points $X_1, \ldots, X_n$, which we assume are drawn independently from the uniform distribution on $S \subset \mathbb{R}^d$, and our goal is to estimate the volume of S. We address this issue from the perspective of the theory of set estimation. In broad terms, set estimation deals with the problem of approximating an unknown set S, or some functionals of S, from sample data consisting of randomly selected points, see [12] for a survey. The problem of estimating the volume of a set has also been widely considered in stereology based on Cavalieri's principle [2, 11, 22]. Stereology consists of mathematical and statistical methods which allow us to provide important quantitative descriptions of the geometry of structures (usually in the three-dimensional space) such as the volume. In contrast to set estimation, data in stereology usually comes in the form of lower dimensional sections of the structure of interest. Regardless the methodology used, the estimation of the volume of a set has practical applications in varied fields such as medicine (organ volume estimation [21], volume estimation of pathology zones from medical image [31]), cell biology (cell volume measurement [28]) or ecology (tree volume estimation [26, 30], home range area or volume estimation [9, 20, 38]).

1.1 Geometric restrictions on S

In our setting, it does not seem feasible to define a unique estimator that efficiently approximate the volume of the unknown set S, unless we restrict the class of sets under consideration. Since the first approaches that assume that S is convex, several proposals

[1] For us a compact domain is a bounded subset which coincides with the closure of its interior.

have dealt with the problem under less restrictive shape conditions. Working on a more general framework than that of convexity allows us to deal with more realistic problems. Before continuing, we introduce some notation. We denote by S^{c}, $\bar{S}$ and ∂S the complement, closure and boundary of S, respectively. Also, we denote by $B(x, r)$ the open ball with center x and radius r. We assume the following:

$$\text{Both } S \text{ and } S^{\mathsf{c}} \text{ satisfy the } r\text{-rolling condition.} \tag{1}$$

Definition 1. A set S is said to fulfill the r-rolling condition if for any $x \in \partial S$, there is a open ball B with radius r such that $B \cap S = \varnothing$ and $x \in \partial B$.

In Figure 1 we show examples of sets that do and do not satisfy condition (1). It is clear that the family of sets that fulfill (1) is much wider than that of convex sets. In particular, it includes sets with holes or inlets without sharp features at the boundary. Our assumption is equivalent to requiring that both S and S^{c} are r-convex [37].

Definition 2. A set S is said to be r-convex if for any point $x \notin \bar{S}$ there is a open ball B of radius r such that $x \in B$ and $B \cap \bar{S} = \varnothing$.

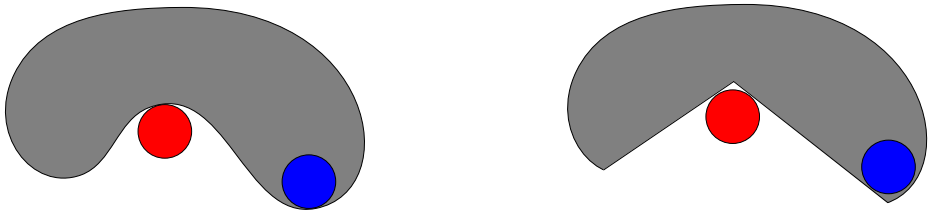

Figure 1: The set S in gray in the left satisfies condition (1). For the set S in gray in the right, neither S nor S^{c} satisfy the r-rolling condition.

We refer to Walther [44] for more details on the rolling condition and its connection to this generalized notion of convexity. Given a set S, its r-convex hull is the smallest

r-convex set that contains S. It is denoted by $C_r(S)$. It can be shown that

$$C_r(S) = (S \oplus rB_1) \ominus rB_1, \tag{2}$$

where B_1 denotes the open ball with center 0 and radius 1, $rA = \{ra : a \in A\}$, $A \oplus C = \{a + c : a \in A, c \in C\}$ and $A \ominus C = \{a : \{a\} \oplus C \subset A\}$. The r-convex hull in (2) can also be written as the intersection of the complements of all the open balls of radius r that do not intersect S. This characterization reminds that of convex sets (with balls instead of half-spaces). The notion of r-convex hull is also closely related to the notion of α-shape, well-known in computational geometry [16]. If, in addition, S is equal to the closure of its interior (which we assume henceforth), then this is also equivalent to asking that ∂S has reach $\geq r$. Following the notation in Federer [17], let $\mathrm{Unp}(S)$ be the set of points $x \in \mathbb{R}^d$ having a unique projection on S

Definition 3. For $x \in S$ let $\mathrm{reach}(S, x) = \sup\{r > 0 : \ B(x, r) \subset \mathrm{Unp}(S)\}$. The reach of S is then defined by $\mathrm{reach}(S) = \inf\{\mathrm{reach}(S, x) : \ x \in S\}$ and S is said to be of positive reach if $\mathrm{reach}(S) > 0$.

Effectively, when ∂S has bounded curvature, the condition is satisfied if $r > 0$ is small enough. We refer to the work by Cuevas et al. [13] for a deep study of the relation between the rolling condition, r-convexity and positive reach.

1.2 Related work on set estimation and our contribution

Plug-in type estimators are a natural choice for the estimation of functionals of S such as the volume. Rényi and Sulanke [39] consider the estimation of the area of a convex set $S \subset \mathbb{R}^2$ with bounded curvature (conditions that imply (1)) using the area of the sample convex hull, obtaining a precise rate of convergence in expectation of order $n^{-2/3}$. Bárány

[5] extends their results to general dimension and Bräker and Hsing [7] to other sampling distributions. One drawback of the plug-in estimator based on the sample convex hull is that it is not rate-optimal for the estimation of the volume. Very recently, Baldin and Reiß [4] reconsider the case of a uniform sampling distribution, but with the added assumption that the sample size is Poisson distributed — in which case the sample comes from a Poisson spatial process with constant intensity over the domain of interest. Under some conditions, they derive the UMVU (uniformly of minimum variance among unbiased estimators) for the volume of a convex set based on a bias correction without sample splitting.

Korostelëv and Tsybakov [27] consider the problem of volume estimation in an image model. One of the settings they assume is that S is of the form $S = \{(x, y) \in [0,1]^2 : y \leq g(x)\}$ for some function g with a given Hölder smoothness. Then the data are of the form $(Z_1, Y_1), \ldots, (Z_n, Y_n)$, with $Z_1, \ldots, Z_n$ IID uniform in $[0,1]^2$ and $Y_i = \xi_i + \mathbb{I}\{Z_i \in S\}$, where the ξ_i's are IID Bernoulli (independent of the Z_i's) and represent the noise. In this setting, they prove a lower bound and provide a rather complex estimator that achieves that lower bound within a poly-logarithmic factor. The class of Hölder smoothness of order 2 is very close to our setting, and for that class Korostelëv and Tsybakov [27] obtain the same error rate as we do here. This work is refined and extended by Gayraud [19], who obtains similar results in arbitrary dimension with unknown sampling distribution. The case of a convex support set is also covered. The underlying method uses sample splitting. In work appearing after ours, Baldin [3] expands on his previous work [4] by considering other classes of sets (including the class considered here) and adopts a sample splitting strategy to correct for the bias.

The work of Gayraud [19], complemented by that of Baldin and Reiß [4], shows that the minimax estimation rate under the assumption of convexity (without smoothness assumption) is $n^{-(d+3)/(2d+2)}$. We prove in Theorem 1 (Section 2) the same minimax estimation

rate under the r-rolling condition (without convexity assumption). Theorem 1 shows, in fact, that adding to the r-rolling condition the assumption of convexity does not make the problem substantially easier from a minimax standpoint.

When the compact domain S is assumed to be r-convex, Rodríguez-Casal [40] proposes to estimate S by taking the r-convex hull of the sample points. This yields a simple plug-in estimator for the volume of S in our setting. However, this plug-in estimator suffers from the drawback of not being rate-optimal. We prove in Theorem 2 (Section 3), that the rate of convergence in expectation is of order $n^{-2/(d+1)}$. Then, we propose in Algorithm 1 an optimal volume estimator based on the sample r-convex hull using a sample splitting strategy. Actually, since r may be unknown in practice, our estimator is based on the sample α-convex hull, where α is a chosen parameter ($\alpha = r$ if r is known). We prove in Theorem 3 that the estimator attains the minimax lower bound. Our method of estimation can be easily enhanced to provide a confidence interval. We briefly discuss this issue in Section 3. Some numerical experiments are presented in Section 4. We discuss some extensions and open problems in Section 5. We give an outline of the proof of Theorem 2 in the Appendix.

2 Minimax lower bound under the rolling condition

Let μ denote the Lebesgue measure of $\mathbb{R}^d$. Also, let $\mathbb{E}_S$ denote the expectation corresponding to $\mathcal{X}_n = \{X_1, \ldots, X_n\}$ sampled IID from the uniform distribution on S. Here we state and proof our result on the lower bound for the volume estimation problem under the r-rolling condition.

Theorem 1. *Let $\mathcal{C}_r(R)$ denote the class of the convex sets S satisfying (1) with half-*

diameter at most R. There is a numerical constant $C > 0$ such that, if $R > 2r$,

$$\inf_{\varphi} \sup_{S \in \mathcal{C}_r(R)} \mathbb{E}_S \left[|\varphi(\mathcal{X}_n) - \mu(S)| \right] \geq C R^2 n^{-(d+3)/(2d+2)}, \tag{3}$$

where the infimum is over all (measurable) functions $\varphi : \mathbb{R}^{dn} \mapsto \mathbb{R}$.

Proof. We employ a simple form of Le Cam's method as expounded in [46, Lem 1]. The construction that follows is similar to that of Mammen and Tsybakov [29] for the problem of set estimation.

Consider the ball centered at the origin of radius R, denoted B_0. Let $y_1, \ldots, y_m$ denote a 2ε-packing of ∂B_0 of maximal size, so that $m \asymp \varepsilon^{-(d-1)}$, as is well-known.[2] The intersection of $\partial B(y_j, \varepsilon)$ and ∂B_0 is the sphere $\partial B(y_j, \varepsilon) \cap H_j$, where $H_j := \{x \in \mathbb{R}^d : \langle x, y_j \rangle = R^2 - \varepsilon^2/2\}$ is a hyperplane. Let $\theta = \operatorname{acos}(1 - \varepsilon^2/2R^2)$, so that 2θ is the aperture of the cone with apex the origin and with base $H_j \cap B_0$. Let C_j denote the corresponding infinite cone. Define another hyperplane $K_j = \{x \in \mathbb{R}^d : \langle x, y_j \rangle = R^2 - Rh\}$, where $h := (R - r)(1 - \cos\theta) = (R - r)\varepsilon^2/2R^2$. Note that H_j is at distance $R - \varepsilon^2/2R$ from the origin, while K_j is parallel to H_j and at distance $R - h$ from the origin. Define the half-space $\bar{K}_j = \{x \in \mathbb{R}^d : \langle x, y_j \rangle \leq R^2 - Rh\}$ and then $\bar{H}_j$ analogously. Let Q_j denote the points $x \in B_0$ with the property that there is a ball B of radius r such that $x \in B \subset B_0 \cap \bar{K}_j$. In other words, we remove from B_0 the cap defined by K_j and obtain Q_j by rolling a ball of radius r inside the resulting set. By construction, $B_0 \cap \bar{H}_j \subset Q_j \subset B_0 \cap \bar{K}_j$, and in particular the different sets $B_0 \cap Q_j^{\mathsf{c}}$, as j varies, do not intersect. (The latter is because $\|y_j - y_{j'}\| > 2\varepsilon$ when $j \neq j'$.) See Figure 2 for an illustration in dimension $d = 2$.

[2] This can be argued as follows. Let vol_{d-1} denote the $(d-1)$-dimensional volume. Note that $\Lambda := \mathrm{vol}_{d-1}(\partial B_0) \asymp 1$. On the one hand, because the sets $U_{j,\varepsilon} := B(y_j, \varepsilon) \cap \partial B_0$ are disjoint, are contained ∂B_0, and have the same $(d-1)$-volume λ_ε, we have $m\lambda_\varepsilon \leq \Lambda$, so that $m\varepsilon^{d-1} = O(1)$. On the other hand, since the sets $U_{j,2\varepsilon}$ cover ∂B_0 and have same $(d-1)$-volume $\lambda_{2\varepsilon}$, we have $m\lambda_{2\varepsilon} \geq \Lambda$. We then conclude with the fact that $\lambda_\varepsilon \asymp \varepsilon^{d-1}$, which can be derived from (6) for example.

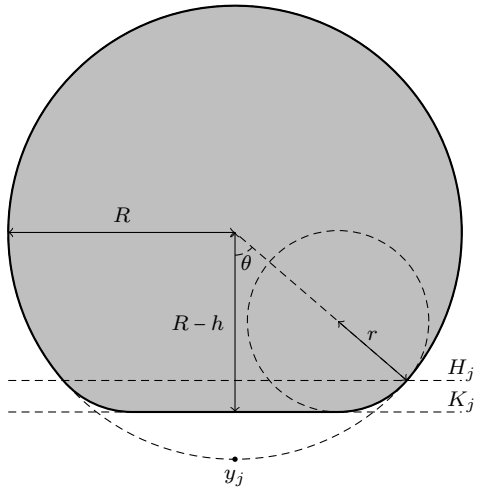

Figure 2: The ball $B_0 = B(0, R)$ is smoothly 'dented' to obtain Q_j, represented in gray.

For $\omega = (\omega_1, \ldots, \omega_m) \in \{0,1\}^m$, let

$$S_\omega = B_0 \cap \bigcap_{\{j:\omega_j=1\}} Q_j. \tag{4}$$

By construction, for any ω, both S_ω^{c} and S_ω satisfy the r-rolling condition, the latter being convex by Lemma 1 in the Appendix.

Let Π_ℓ denote the uniform distribution on $\Omega := \{\omega : |\omega|_1 = \ell\}$, where for $\omega = (\omega_1, \ldots, \omega_m) \in \{0,1\}^m$, we let $|\omega|_1 = \sum_j \omega_j$. The parameter ℓ will be chosen later on. Define $\eta = \mu(B_0) - \mu(Q_j) = \mu(B_0 \setminus Q_j)$. By [46, Lem 1],

$$\inf_{\varphi} \sup_{S \in \mathcal{C}_r(R)} \mathbb{E}_S \, |\varphi(\mathcal{X}_n) - \mu(S)| \geq \tfrac{1}{2}\ell\eta \Big(1 - \tfrac{1}{2}\mathrm{TV}(P_0^{\otimes n}, P_1^{\otimes n})\Big), \tag{5}$$

where P_0 is the uniform distribution on B_0, P_1 is the mixture of $\mathbb{P}_{S_\omega}$ when $\omega \sim \Pi_\ell$, and TV denotes the total variation metric for distributions. This is the bound we work with.

We first bound η, from below but also from above, as this will be needed later on. Let γ

denote the angle associated to K_j as θ is associated to H_j, and note that $\gamma = \operatorname{acos}(1-h^2/R^2)$. We will take ε small, and as $\varepsilon \to 0$ we have that $m \to \infty$, $\theta \sim \varepsilon/R$, and $\gamma \sim \sqrt{2h/R}$.

The volume of a cap of the unit ball in $\mathbb{R}^d$ at distance $1-t$ from the origin is equal to

$$\frac{\pi^{\frac{d-1}{2}}}{\Gamma\left(\frac{d+1}{2}\right)} \int_0^{\arccos(1-t)} \sin^d(x)\mathrm{d}x \asymp t^{(d+1)/2}, \quad t \to 0. \tag{6}$$

Using this, as $\varepsilon \to 0$,

$$\begin{aligned} &\eta \le \mu(B_0 \setminus \bar{H}_1) \asymp \varepsilon^{d+1}, \\ &\eta \ge \mu(B_0 \setminus \bar{K}_1) \asymp h^{(d+1)/2} \asymp \varepsilon^{d+1}, \quad \text{since } h \asymp \varepsilon^2. \end{aligned} \tag{7}$$

Define

$$Z = \frac{\mathrm{d}P_1^{\otimes n}(\mathcal{X}_n)}{\mathrm{d}P_0^{\otimes n}(\mathcal{X}_n)} = (1-\ell\eta/\zeta_d R^d)^{-n} \, \frac{1}{|\Omega|} \sum_{\omega \in \Omega} \mathbb{I}\{\mathcal{X}_n \subset S_\omega\}, \tag{8}$$

where ζ_d is the volume of the unit ball in dimension d. Then[3]

$$\mathrm{TV}(P_0^{\otimes n}, P_1^{\otimes n}) = \mathbb{E}_0[|Z-1|] \le \sqrt{\mathbb{E}_0(Z^2)-1}, \tag{9}$$

where the inequality is Cauchy-Schwarz's. We have

$$\mathbb{E}_0(Z^2) = (1-\ell\eta/\zeta_d R^d)^{-2n} \, \frac{1}{|\Omega|^2} \sum_{\omega,\omega' \in \Omega} \mathbb{E}_0(\mathbb{I}\{\mathcal{X}_n \subset S_\omega\}\mathbb{I}\{\mathcal{X}_n \subset S_{\omega'}\}), \tag{10}$$

with

$$\mathbb{E}_0(\mathbb{I}\{\mathcal{X}_n \subset S_\omega\}\mathbb{I}\{\mathcal{X}_n \subset S_{\omega'}\}) = \mathbb{E}_0(\mathbb{I}\{\mathcal{X}_n \subset S_\omega \cap S_{\omega'}\}) = \left(1-(2\ell - |\omega \wedge \omega'|_1)\eta/\zeta_d R^d\right)^n, \tag{11}$$

where $\omega \wedge \omega' = (\omega_1 \wedge \omega_1', \ldots, \omega_m \wedge \omega_m')$ when $\omega = (\omega_1, \ldots, \omega_m)$ and $\omega' = (\omega_1', \ldots, \omega_m')$. Noting that $|\omega \wedge \omega'|_1$ has the hypergeometric distribution with parameters (m, ℓ, ℓ) when ω, ω' are

[3]The quantity on the right-hand side of this inequality is the square-root of the chi-squared divergence [42, Page 86].

IID with distribution Π_ℓ, and letting V denote a random variable with that distribution, we have

$$\begin{aligned}
\mathbb{E}_0(Z^2) &= (1-\ell\eta/\zeta_d R^d)^{-2n}\,\mathbb{E}\Big[\big(1-(2\ell-V)\eta/\zeta_d R^d\big)^n\Big] \\
&= \mathbb{E}\left[\left(\frac{1-(2\ell-V)\eta/\zeta_d R^d}{(1-\ell\eta/\zeta_d R^d)^2}\right)^n\right] \\
&\le \mathbb{E}\Big[\big(1+V\eta/\zeta_d R^d + 10(\ell\eta/\zeta_d R^d)^2\big)^n\Big] \\
&\le \exp(10(\ell\eta/\zeta_d R^d)^2 n)\,\mathbb{E}\big[\exp(nV\eta/\zeta_d R^d)\big],
\end{aligned} \tag{12}$$

where in the third line we assumed that $\ell\eta/\zeta_d R^d \le 1/2$. The function $x \mapsto e^{ax}$ (with $a > 0$ fixed) being convex, we may apply [23, Th 4] to bound the last expectation by

$$\mathbb{E}\big[\exp(nW\eta/\zeta_d R^d)\big], \tag{13}$$

where W is binomial with parameters $(\ell, \ell/m)$. We then continue

$$\mathbb{E}\big[\exp(nW\eta/\zeta_d R^d)\big] = \Big(1-\tfrac{\ell}{m}+\tfrac{\ell}{m}e^{n\eta/\zeta_d R^d}\Big)^\ell \le \exp\Big[\tfrac{\ell^2}{m}e^{n\eta/\zeta_d R^d}\Big]. \tag{14}$$

Therefore, we conclude that $\mathbb{E}_0(Z^2) \le 2$ when

$$\exp(10(\ell\eta/\zeta_d R^d)^2 n) \le \sqrt{2}, \tag{15}$$

(which implies $\ell\eta/\zeta_d R^d \le 1/2$) and when

$$\exp\big[\tfrac{\ell^2}{m}e^{n\eta/\zeta_d R^d}\big] \le \sqrt{2}. \tag{16}$$

From (7), we know there is a constant c_0 such that $\eta/\zeta_d R^d \le c_0\varepsilon^{d+1}$. Hence (15) and (16) are implied by

$$\ell^2 \le \tfrac{1}{10c_0^2}\log(\sqrt{2})/n\varepsilon^{2d+2}, \quad \ell^2 \le \varepsilon^{1-d}e^{-c_0 n\varepsilon^{d+1}}\log(\sqrt{2}). \tag{17}$$

Taking $\varepsilon = n^{-1/(d+1)}$, we can see that we may set $\ell = \lceil cn^{(d-1)/(2d+2)} \rceil$ with $c > 0$ a sufficiently small constant. Note that $\eta \asymp 1/n$ with this choice of ε by (7). This guarantees that, n being large enough, $\mathbb{E}_0(Z^2) \leq 2$, and when this is the case, from (9), the RHS of (5) is lower-bounded by $\eta\ell/4 \asymp (1/n)n^{(d-1)/(2d+2)} = n^{-(d+3)/(2d+2)}$, which concludes the proof of Theorem 1. □

3 Plug-in estimator and bias correction

In this section we first consider the estimation of the volume of S using the volume of the r-convex hull of the sample. We prove that this estimator is not rate optimal. Then in Section 3.2 we propose a data splitting approach to correct the bias of the plug-in estimator and prove that it is rate optimal. In Section 3.3 we discuss the construction of confidence intervals for the volume.

3.1 Plug-in estimator based on the r-convex hull

Under the r-rolling condition, our first attempt to estimate the volume of S is to consider the volume of the r-convex hull of the sample points. In fact, since r may be unknown, we consider the sample α-convex hull, $C_\alpha(\mathcal{X}_n)$, where α is a chosen parameter ($\alpha = r$ if r is known). We state in Theorem 2 below that this plug-in estimator does not achieve the minimax rate. Theorem 2 generalizes [36, Thm 1] to the d-dimensional Euclidean space. A skecth of the proof can be found in the Appendix. Although some arguments are analogous to those used in the bidimensional case, the proof of Theorem 2 is not just an extension of that for [36, Thm 1]. In particular, see Lemmas 2 and 3 in the Appendix.

Theorem 2. *Assume that S satisfies (1) and the half-diameter of S is at most R. Let $\mathcal{X}_n = \{X_1, \ldots, X_n\}$ be a random sample from the uniform distribution on S. Given $\alpha \in (0, r]$,*

there exits a constant c which only depends on (d, r, R, α) such that, for all n,

$$\mathbb{E}_S\left[\mu(C_\alpha(\mathcal{X}_n) \triangle S)\right] \leq cn^{-2/(d+1)}, \tag{18}$$

where $\triangle$ denotes the symmetric difference and also

$$\mathbb{P}(\mu(C_\alpha(\mathcal{X}_n) \triangle S) > \varepsilon) \leq c\varepsilon^{-d}\exp(-n\varepsilon^{(d+1)/2}/c), \quad \forall \varepsilon > 0. \tag{19}$$

Remark 1. Assuming that $\alpha \leq r$ we have that

$$\mu(C_\alpha(\mathcal{X}_n) \triangle S) = \mu(S) - \mu(C_\alpha(\mathcal{X}_n)), \tag{20}$$

so that, by (18), the plug-in estimator $\mu(C_\alpha(\mathcal{X}_n))$ achieves the error rate $O(n^{-2/(d+1)})$. We conjecture that this is sharp, and if so, the plug-in estimator does not achieve the error rate obtained in Theorem 1, not even within a poly-logarithmic factor.

Remark 2. The results of Theorem 2 can be extended to cover a more general sampling distribution P_X, and α equal to α_n, which may depend on the sample size, but not on the sample. Assuming that the probability distribution P_X satisfies that there exists $\delta > 0$ such that $P_X(C) \geq \delta\mu(C \cap S)$ for all Borel subsets $C \subset \mathbb{R}^d$, and the sequence $\{\alpha_n\}$ satisfies

$$\lim_{n\to\infty} \frac{n\alpha_n^d}{\log n} \to \infty, \tag{21}$$

it can be shown, using similar arguments as in Theorem 2, that

$$\mathbb{E}\left[\mu(C_{\alpha_n}(\mathcal{X}_n) \triangle S)\right] = O\left(\alpha_n^{-(d-1)/(d+1)} n^{-2/(d+1)}\right). \tag{22}$$

3.2 Minimax optimal estimator

In order to correct the bias in the plug-in estimator, we propose a sample splitting strategy. Details are provided in Algorithm 1. We proof in Theorem 3 that the proposed estimator

achieves the minimax rate over the class of sets that satisfy (1) (and are not necessarily convex).

Algorithm 1 Volume estimation with the sample α-convex hull and bias correction

0 – Input: Sample $\mathcal{X}_n$, size of the first subsample m, $\alpha > 0$.

1 – Sample splitting: Randomly split the sample $\mathcal{X}_n$ into two subsamples $\mathcal{X}'_n$ and $\mathcal{X}''_n$ of respective sizes m and $n - m$.

2 – Set estimation: Compute the α-convex hull of the first subsample to get an estimate for the support set S,

$$\hat{S} := C_\alpha(\mathcal{X}'_n). \tag{23}$$

3 – Bias estimation: Compute the proportion of points in the second subsample that fall outside $\hat{S}$,

$$\hat{p} := \frac{\#(\mathcal{X}''_n \setminus \hat{S})}{n - m}. \tag{24}$$

4 – Output: Return the estimator

$$\hat{V} = \frac{\mu(\hat{S})}{(1 - \hat{p}) \vee 1/2}. \tag{25}$$

The practical implementation of the proposed estimator requires the computation of the α-convex hull and its volume. The algorithm for computing the α-convex hull is described in [15] and is implemented in the two-dimensional case in the R-package *alphahull* [35].

Theorem 3. *Assume that S satisfies (1) and the half-diamenter of S is at most R. Fix $\alpha \in (0, r]$ and $\beta \in (0, 1/2)$, and take m such that $\beta \le m/n \le 1 - \beta$. Then estimator $\hat{V}$ defined in Algorithm 1 satisfies*

$$\mathbb{E}_S\left[|\hat{V} - \mu(S)|\right] \le c\, n^{-(d+3)/(2d+2)}, \tag{26}$$

for some $c > 0$ that only depends on (d, r, R, α, β).

Proof. Below $c_1, c_2, \ldots$ denote quantities that only depend on (d, r, R, α, β). In that regard, we will use the fact that, when S has half-diameter bounded by R, then $\mu(S) \le \mu(B(0,R)) = \omega_d R^d$, where ω_d is the volume of the unit ball in $\mathbb{R}^d$.

Recall the process leading to $\hat{V}$ in Algorithm 1. Define $\tilde{p} = \mu(S \setminus \hat{S})/\mu(S)$, as well as $V = \mu(S)$, which is what we want to estimate, and $\hat{V}_0 = \mu(\hat{S})$, which is the plug-in estimate. We have

$$\begin{aligned}\mathbb{E}_S |\hat{V} - V| = \mathbb{E}_S \big[|\hat{V} - V| \mathbb{I}\{\tilde{p} \ge 1/4\}\big] \\ + \mathbb{E}_S \big[|\hat{V} - V| \mathbb{I}\{\tilde{p} < 1/4, \hat{p} > 1/2\}\big] \\ + \mathbb{E}_S \big[|\hat{V} - V| \mathbb{I}\{\tilde{p} < 1/4, \hat{p} \le 1/2\}\big].\end{aligned} \tag{27}$$

By (25), we have $\hat{V} \le 2\hat{V}_0 \le 2V$, and by the definition of $\tilde{p}$ and the fact that $\hat{S} \subset S$ by construction of $\hat{S}$, we also have $\mu(S \setminus \hat{S}) = \mu(S) - \mu(\hat{S})$, so that $\tilde{p} = (V - \hat{V}_0)/V \le \mu(\hat{S} \triangle S)/V$. Using these bounds, we derive

$$\begin{aligned}\mathbb{E}_S \big[|\hat{V} - V| \mathbb{I}\{\tilde{p} \ge 1/4\}\big] &\le V \, \mathbb{P}\big(\tilde{p} \ge 1/4\big) \\ &\le V \, \mathbb{P}\big(\mu(\hat{S} \triangle S) \ge V/4\big) \\ &\le c_1 \exp(-n/c_1),\end{aligned} \tag{28}$$

by (19) and the fact that $V \le \omega_d R^d$, and also the fact that $m \ge \beta n$.

Similarly, we have

$$\begin{aligned}\mathbb{E}_S \big[|\hat{V} - V| \mathbb{I}\{\tilde{p} < 1/4, \hat{p} > 1/2\}\big] &\le V \, \mathbb{P}\big(\tilde{p} < 1/4, \hat{p} > 1/2\big) \\ &\le V \, \mathbb{P}\big(\hat{p} > 1/2 \,\big|\, \tilde{p} < 1/4\big) \\ &\le c_2 \exp(-n/c_2),\end{aligned} \tag{29}$$

using the fact that, given $\tilde{p}$, $(n-m)\hat{p} \sim \mathrm{Bin}(n-m, \tilde{p})$, which is stochastically bounded by $\mathrm{Bin}(n-m, 1/4)$ when $\tilde{p} < 1/4$, and then applying Bernstein's inequality together with the fact that $n - m \ge \beta n$.

Finally, because

$$|\hat{V} - V| = \hat{V}_0 \frac{|\hat{p} - \tilde{p}|}{(1-\hat{p})(1-\tilde{p})} \leq \tfrac{8}{3} V |\hat{p} - \tilde{p}|, \tag{30}$$

when $\tilde{p} < 1/4$ and $\hat{p} \leq 1/2$, we have

$$\begin{aligned}
\mathbb{E}_S\left[|\hat{V} - V| \, \mathbb{I}\{\tilde{p} < 1/4, \hat{p} \leq 1/2\}\right] &\leq \tfrac{8}{3} V \, \mathbb{E}_S\left[|\hat{p} - \tilde{p}|\right] \\
&\leq \tfrac{8}{3} V \sqrt{\mathbb{E}_S\left[(\hat{p} - \tilde{p})^2\right]} \\
&= \tfrac{8}{3} V \sqrt{\mathbb{E}_S\left[\tilde{p}(1-\tilde{p})/(n-m)\right]} \\
&\leq \tfrac{8}{3} V \sqrt{\mathbb{E}_S\left[\tilde{p}/(\beta n)\right]} \\
&= c_3 V \sqrt{\mathbb{E}_S\left[\mu(\hat{S} \vartriangle S)\right]/n} \\
&\leq c_4 \sqrt{(1/n)^{2/(d+1)}/n} = c_4 n^{-(d+3)/(2d+2)},
\end{aligned} \tag{31}$$

using the Cauchy-Schwarz inequality, the fact that, given $\tilde{p}$, $(n-m)\hat{p} \sim \mathrm{Bin}(n-m, \tilde{p})$, with $n - m \geq \beta n$, and (18).

Combining all bounds, and noticing that $(d+3)/(2d+2) \leq 1$ for all $d \geq 1$, proves the result. □

Remark 3. The estimator we propose in Algorithm 1, just like the plug-in estimator, depends on the choice of $\alpha > 0$. We proved our result (Theorem 3) under the assumption that $\alpha \leq r$, but in general r is unknown. Also, although the theory works for any α thus chosen, in practice, an optimal choice for α may depend on the sample size. Under uniform sampling, Rodríguez-Casal and Saavedra-Nieves [41] propose a data-driven selector of α, α_n, such that, with probability one, satisfies $\alpha_n \leq r$ and $\alpha_n \to r$.

Remark 4. This estimator relies heavily on the fact that the sampling distribution is uniform. If this is not the case, it can be biased downward or upward. For example, suppose

that S is the unit disc in dimension $d = 2$ and that the sampling distribution has the following density

$$f(x) = \frac{a\mathbb{I}\{\|x\| \le 1/2\} + b\mathbb{I}\{1/2 < \|x\| \le 1\}}{\frac{\pi}{4}a + \frac{3\pi}{4}b}, \tag{32}$$

where $a, b > 0$. In that case, with probability approaching 1 as $m \to \infty$,

$$\tilde{p} = c\mu(S \setminus C_\alpha(\mathcal{X}'_n))/\mu(S), \quad c := \frac{b}{a/4 + 3b/4}, \tag{33}$$

and by varying a and b, c can be any real in $(0, 4/3]$. (Indeed, due to Theorem 2, $S \setminus \hat{S} \subset B(\partial S, 1/2)$ with probability tending to 1, and conditional on $\mathcal{X}'_n$ satisfying this condition, the probability that a point from $\mathcal{X}''_n$ falls in $S \setminus \hat{S}$ is equal to $\int_{S\setminus\hat{S}} f(x)\mathrm{d}x = (c/\pi)\mu(S \setminus \hat{S})$ since $f(x) = c/\pi$ for all $x \in S \cap B(\partial S, 1/2)$.) As a consequence, if $c < 1$, meaning $a > b$, the estimator $\hat{V}$ remains biased downward, while if $c > 1$, meaning $a < b$, it is biased upward. In both cases the bias correction fails and the estimator $\hat{V}$ can only be shown to achieve the rate of the plug-in estimator of Remark 1.

3.3 Confidence interval for the volume

Beyond a point estimate, the procedure can be modified to yield a confidence interval. For $\eta \in (0, 1)$, $k > 0$ integer, and $z \in \{0, \dots, k\}$, let $A_{\eta,k}(z)$ and $B_{\eta,k}(z)$ be such that, when $Z \sim \mathrm{Bin}(k, \theta)$,

$$\mathbb{P}\left(A_{\eta,k}(Z) \le \theta \le B_{\eta,k}(Z)\right) \ge 1 - \eta. \tag{34}$$

Thus, $[A_{\eta,k}(Z), B_{\eta,k}(Z)]$ is a $(1-\eta)$-level confidence interval for θ. The construction we have in mind is the 'exact' confidence interval of Clopper and Pearson [10]. For other constructions, such as the basic interval based on the normal approximation and the refinement due to Wilson [45], the inequality is only approximate — see [8] for a recent discussion. In Algorithm 2 we describe how to generate a confidence interval for the volume based on a particular confidence interval for a binomial proportion.

Algorithm 2 Confidence interval for the volume based on the sample α-convex hull

0 – Input: Sample $\mathcal{X}_n$, size of the first subsample m, $\alpha > 0$, confidence level $1-\eta$

1 – Estimation: split the sample and compute $\hat{S}$ and $\hat{p}$ as in Algorithm 1.

2 – Output: Return the interval

$$\left[\frac{\mu(\hat{S})}{1-A_{\eta,n-m}((n-m)\hat{p})}, \frac{\mu(\hat{S})}{1-B_{\eta,n-m}((n-m)\hat{p})}\right]. \tag{35}$$

Proposition 1. *In the setting of Theorem 3, and assuming that (34) holds, $\mu(S)$ is contained in the interval (35) with probability at least $1-\eta$.*

Proof. We keep the notation introduced in Section 3.2. Define $k = n-m$ and $Z = (n-m)\hat{p}$. The result hinges on the fact that $Z \sim \mathrm{Bin}(k, \tilde{p})$ when conditioning on $\tilde{p}$, which with (34) implies that

$$A_{\eta,k}(Z) \le \tilde{p} \le B_{\eta,k}(Z), \tag{36}$$

with probability at least $1-\eta$. When this is the case, we 'pivot' on $\tilde{p} = (V - \hat{V}_0)/V$ to obtain

$$\frac{\hat{V}_0}{1-A_{\eta,k}(Z)} \le V \le \frac{\hat{V}_0}{1-B_{\eta,k}(Z)}. \tag{37}$$

□

Remark 5 (Length of the confidence interval). We believe similar analysis carries out more broadly, but for concreteness and simplicity consider the basic construction based on the normal approximation where

$$A_{\eta,k}(z) = \frac{z}{k} - q_\eta\sqrt{\frac{\frac{z}{k}(1-\frac{z}{k})}{k}}, \quad B_{\eta,k}(z) = \frac{z}{k} + q_\eta\sqrt{\frac{\frac{z}{k}(1-\frac{z}{k})}{k}}, \tag{38}$$

where q_η is the $1-\eta/2$ quantile of the standard normal distribution. Although the inequality in (34) is only approximate, it becomes an equality when $\theta = \theta_k$ is such that $k(\theta_k \wedge (1 -$

$\theta_k)) \to \infty$. We now apply this to our context. Although this is a conjecture, we have reasons to believe that (18) is essentially sharp in that $\mu(C_\alpha(\mathcal{X}_n) \triangle S) \asymp n^{-2/(d+1)}$ under our conditions. (This statement and those that follow are to be understood in probability.) Assuming this is the case, and using the notation introduced in Section 3.2, we have $\tilde{p} = (V - \hat{V}_0)/V \asymp m^{-2/(d+1)}$, and the interval is asymptotically accurate in level when $(n-m)\tilde{p} \asymp (n-m)m^{-2/(d+1)} \to \infty$. From now on, we take $m \geq n/2$, so that this holds. Then, letting A and B be short for $A_{\eta,n-m}((n-m)\hat{p})$ and $B_{\eta,n-m}((n-m)\hat{p})$ respectively, the interval (35) is of length

$$\frac{\hat{V}_0(B-A)}{(1-B)(1-A)}, \tag{39}$$

with $\hat{V}_0 \to V$, while $A \sim \tilde{p} = o(1)$, $B \sim \tilde{p} = o(1)$, and $B - A = 2q_\eta\sqrt{\hat{p}/(n-m)} = O(\sqrt{\tilde{p}/(n-m)} = O(m^{-1/(d+1)}n^{-1/2})$. Thus the length of the interval is of order $O(n^{-(d+3)/(2d+2)})$ when taking $m \asymp n$.

4 Numerical experiments

Here we discuss the results of a simulation study that illustrates the performance of the proposed volume estimator $\hat{V}$ defined in Algorithm 1. We consider the set

$$S = \{x \in \mathbb{R}^2 : \ 0.25 \leq \|x\| \leq 0.5\}. \tag{40}$$

Note that S is r-convex for $r = 0.25$ and $\mu(S) = \pi(0.5^2 - 0.25^2)$.

In the first experiment, we generate a sample of size n from the uniform distribution on S and calculate the estimator $\hat{V}$. We consider different sizes m for the subsample $\mathcal{X}'_n$ (different ways of splitting the sample) and different values of α. Each setting is repeated $B = 500$ times. Figure 3 shows the mean values $|\mu(S) - \hat{V}|/\mu(S)$ over the B repeats (error bars represent one standard deviation). We do the same for $\mu(C_\alpha(\mathcal{X}_n))$ instead of $\hat{V}$.

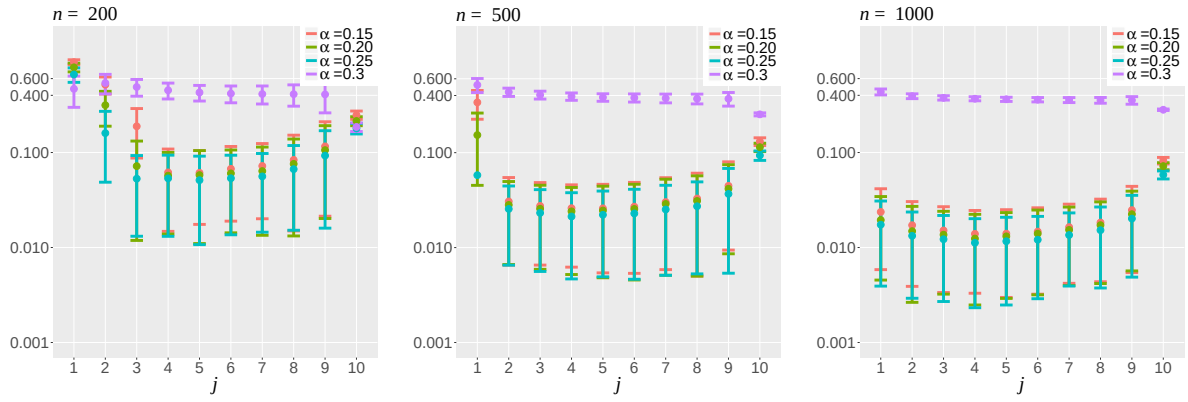

Figure 3: Mean values of $|\mu(S) - \hat{V}|/\mu(S)$ over $B = 500$ repeats and error bars representing one standard deviation (a base-10 log scale is used for the Y-axis). In the computation of $\hat{V}$ we have considered different sizes $m = [nj/10]$ for the subsample $\mathcal{X}'_n$. The case $j = 10$ corresponds to the plug-in estimator.

We observe that our proposal for the estimation of the volume considerably reduces the bias of $\mu(C_\alpha(\mathcal{X}_n))$ (case $j = 10$). The value of j determines the proportion of sample for estimating S. Small values of j result in a poor estimation of the set so that the volume correction does not work, and the other way round for large values of j. More research is needed in order to determine if there is any optimality in the choice of the parameter β. Regarding the choice of the parameter α, our method works for $\alpha \leq r$, as expected. For larger values of α, the estimation is very poor. Note that, when $\alpha > r$, $C_\alpha(\mathcal{X}_n)$ tends to fill in the "hole" of S so that the estimated volume will be increased, even if no bias correction is made.

Bagging In the second experiment, we study the same scenarios as in the first experiment, and examine the strategy of performing the sample splitting multiple times. The new estimator, denoted $\hat{V}_{\text{bag}}$, is obtained by computing $\hat{V}$ for $b = 100$ random sample splittings

and averaging these. This is a form of bagging [18, Sec 8.7], therefore the name. Figure 4 shows the mean values of $|\mu(S) - \hat{V}_{\text{bag}}|/\mu(S)$ over the B repeats. ($B = 500$ as before.) As it can be seen, compared to Figure 3, this bagging technique reduces the variance of the error.

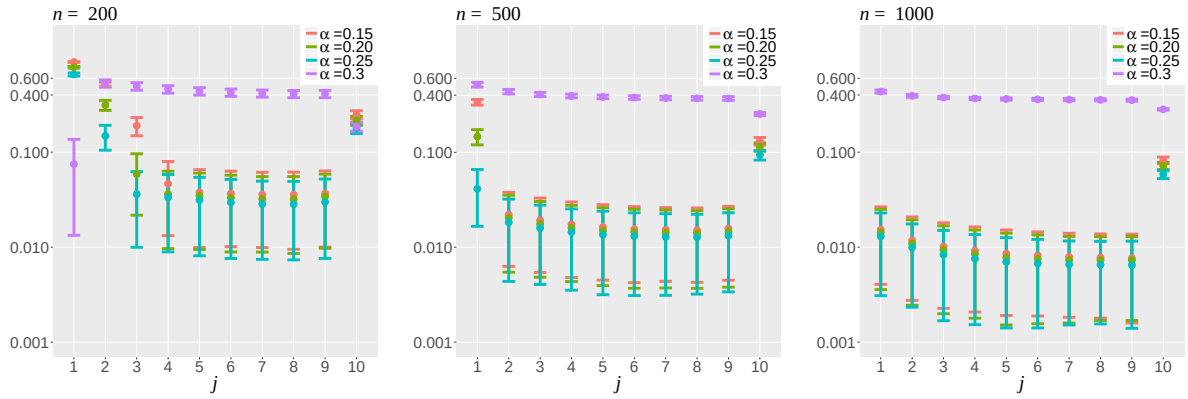

Figure 4: Mean values of $|\mu(S) - \hat{V}_{\text{bag}}|/\mu(S)$ over $B = 500$ repeats and error bars representing one standard deviation (a base-10 log scale is used for the Y-axis). The bagging was over $b = 100$ sample splittings. In the computation of $\hat{V}$ we have considered different sizes $m = [nj/10]$ for the subsample $\mathcal{X}'_n$. The case $j = 10$ corresponds to $\hat{V} = \mu(C_\alpha(\mathcal{X}_n))$.

Confidence intervals As mentioned before, the proposed method lends itself naturally to the computation of confidence intervals for $\mu(S)$ based on the computation of confidence intervals for $\tilde{p}$. We use the method of Wilson [45] for that purpose. Results of the estimated coverage probability and estimated mean length of the confidence intervals for different nominal confidence levels are shown in Table 1. (This is just meant as a proof of concept since there are no other methods we know off to compare this with.)

	Level	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95
$n = 200$	Coverage	0.496	0.540	0.614	0.664	0.704	0.754	0.790	0.846	0.894	0.954
	Length	0.051	0.058	0.064	0.071	0.079	0.088	0.098	0.111	0.127	0.152
$n = 500$	Coverage	0.470	0.544	0.602	0.648	0.692	0.728	0.786	0.842	0.896	0.944
	Length	0.021	0.024	0.027	0.030	0.033	0.036	0.041	0.046	0.052	0.063
$n = 1000$	Coverage	0.502	0.546	0.594	0.652	0.718	0.756	0.804	0.852	0.878	0.938
	Length	0.012	0.013	0.014	0.016	0.018	0.020	0.022	0.025	0.028	0.034

Table 1: Coverage and length of the confidence interval for $\mu(S)$ based on a confidence interval for $\tilde{p}$. We split the sample in half (meaning, we used $m = n/2$) and used $\alpha = 0.25$. Each setting is repeated $B = 500$ times and what are shown are the averages of the B repeats.

The convex case We replicated the study in [4] to compare the performance of our estimator $\hat{V}$ with that of the estimators discussed in that paper for the convex case. Data points are simulated for an ellipse, S, with center at the origin, major axis of length 10 and minor axis of length 4; see Figure 5. More specifically, for different values of n, we generated $B = 500$ samples from a Poisson spatial process over S with constant intensity $\lambda = n/\mu(S)$. The size of each sample, N, is Poisson distributed with mean n. For the computation of $\hat{V}$ we randomly split each sample into two subsamples of equal size and compute the α-convex hull of the first subsample with $\alpha = 10$. We base our choice of α on the fact that, under the assumption of convexity, the α-convex hull estimator works reasonably well for large values of α. Figure 6 (right) shows, for the considered estimators, the RMSE normalized by the true area based on the $B = 500$ Monte Carlo iterations for each n. We use the same notation as in [4]. Our estimator $\hat{V}$ performs slightly better than the rate-optimal estimator based on sample splitting by Gayraud [19], denoted by $\hat{v}_G$. The

best performance corresponds to $\hat{v}_{oracle}$, altough its computation depends on the unknown intensity λ. The estimators $\hat{v}_{plugin}$ and $\hat{v}$, for the case of unknown intensity λ, also perform well. As already pointed out by Baldin and Reiß [4], all these methods clearly outperform the results of other estimators that are not rate-optimal, such as the Lebesgue measure of the convex hull of the sample, denoted by $|\hat{C}|$, and the so-called naive oracle estimator N/λ. Comparisons are easier to make in Figure 6 (right), where we report relative errors, e.g. quotient with respect to the oracle estimator.

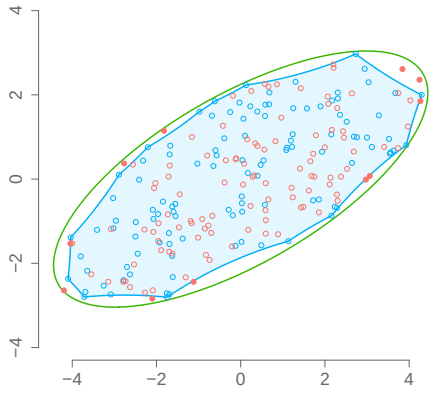

Figure 5: A random sample is generated on the ellipse S in green and splitted into $\mathcal{X}_n'$ (blue points) and $\mathcal{X}_n''$ (open and solid red points). The solid red points are the observations of $\mathcal{X}_n''$ that fall outside $C_\alpha(\mathcal{X}_n')$, represented in blue for $\alpha = 10$.

5 Extensions

Piecewise smooth boundary. We are confident that our proof arguments proceed with relatively minor modifications when ∂S is piecewise smooth. However, our working condition

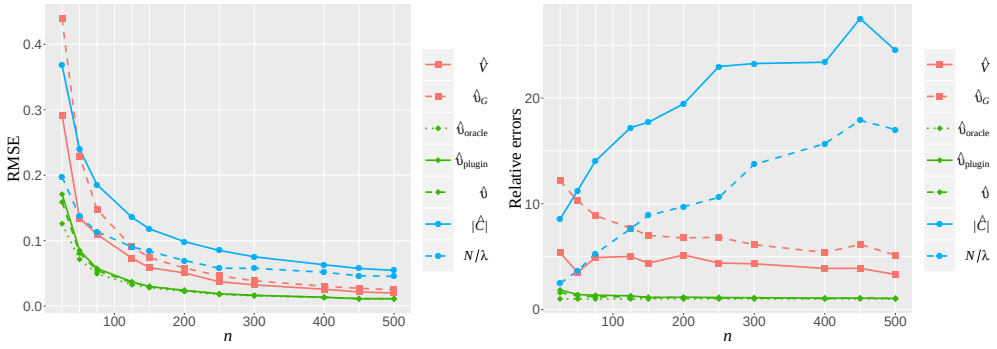

Figure 6: Left: The RMSE (normalized by $\mu(S)$) of different estimators of $\mu(S)$, based on $B = 500$ Monte Carlo iterations in each case. Right: relative errors (quotient with respect to the oracle estimator).

(1) — which is equivalently expressed as a requirement on the reach of ∂S — is simple and compact, and the resulting analysis already contains all the intricacies.

Non-uniform sampling. A more substantial extension is to the setting of an unknown sampling distribution. This setting is considered in [19], where the sampling density is estimated by a standard kernel procedure, and that estimate is incorporated in the estimator of the volume. Although the methodology and theory developed in that paper do not apply directly, an adaptation to our setting (namely, to sets satisfying (1)) seems viable.

Perimeter estimation. A parallel line of research tackles the problem of estimating the perimeter of S. In fact, this problem is also considered by Rényi and Sulanke [39] and Bräker and Hsing [7], still in the context of a convex support set in dimension $d = 2$. More recently, in the same setting as ours here, but restricted to dimension $d = 2$, Cuevas et al. [13] study the perimeter of the sample α-convex hull, while Arias-Castro and Rodríguez-Casal [1] study the perimeter of the sample α-shape. Parallel to the work of Korostelëv and Tsybakov [27], and working with a similar model, we find the work of Kim and Korostelev

[25]. We also mention a series of papers that consider the closely related problem of estimating the Minkowski content of the boundary of S, still under a similar model, making various regularity assumptions on S [14, 24, 33, 34]. It would be interesting to obtain similar results for the problem of estimating the perimeter of S under our setting or some of these other settings.

A Appendix: additional proofs

A.1 Rolling a ball inside a convex set

Definition 4. For a set S and $\alpha > 0$, let $G_\alpha(S)$ denote the set of $x \in S$ with the property that there is an open ball B of radius α such that $x \in B \subset S$.

Lemma 1. *If $S \subset \mathbb{R}^d$ is convex, then for any $\alpha > 0$, $G_\alpha(S)$ is either empty or convex.*

Proof. Write G for $G_\alpha(S)$. By definition (and the axiom of choice), for any $x \in S$ we may choose an open ball of radius α, denoted B_x, such that $x \in B_x \subset S$. Suppose S contains a ball of radius α, for otherwise G is empty and there is nothing else to prove. Take $x, y \in G$ and let C denote the convex hull of $B_x \cup B_y$. On the one hand, $C \subset S$, because $B_x \cup B_y \subset S$ and S is convex. On the other hand, for all $z \in C$, there is a ball B of radius α such that $z \in B \subset C$. This is obvious from the fact that C is the union of the cylinder with axis $[xy]$ and radius α and the two half balls defined by B_x and B_y on each end, which can be expressed as

$$C = \bigcup_{t \in [x,y]} B(t, \alpha). \tag{41}$$

Hence, $C \subset G$, and in particular, $[xy] \subset G$ since $[xy] \subset C$. This being true for all $x, y \in G$, we conclude that G is indeed convex. □

A.2 Proof of Theorem 2

In this section we sketch the proof of Theorem 2. More details can be found in [32]. First, we need to introduce some notation and state two auxiliary results. The distance between $x \in \mathbb{R}^d$ and $S \subset \mathbb{R}^d$ is defined as $\operatorname{dist}(x, S) = \inf_{s \in S} \|x - s\|$. Given a unit vector u and an angle $\theta \in [0, \pi/2]$, consider the infinite cone with apex x, axis u and aperture 2θ defined by

$$C_u^\theta(x) = \left\{ z \in \mathbb{R}^d, z \neq x : \langle z - x, u \rangle \geq \|z - x\| \cos\theta \right\}.$$

For $h > 0$, consider the *finite* cone obtained by intersecting an infinite cone with a ball of radius h centered at its apex, $C_{u,h}^\theta(x) = C_u^\theta(x) \cap B(x, h)$. For $x \in \mathbb{R}^d$ and $r > 0$, let $\mathcal{E}_{x,r} = \{B(y, r) : y \in B(x, r)\}$.

Definition 5. The family of sets $\mathcal{U}$ is said to be unavoidable for another family of sets $\mathcal{E}$ if, for all $E \in \mathcal{E}$, there exists $U \in \mathcal{U}$ such that $U \subset E$.

Lemma 2. *Under the conditions of Theorem 2, for any $x \in S$ such that $\operatorname{dist}(x, \partial S) > \alpha/2$, with $0 < \alpha \leq r$, there exists a family $\mathcal{U}_{x,\alpha}$ with at most m_1 elements that is unavoidable for $\mathcal{E}_{x,\alpha}$ and satisfies*

$$\mu(U \cap S) \geq c_1 \alpha^d, \quad \forall U \in \mathcal{U}_{x,\alpha},$$

where $c_1, m_1 \geq 1$ depend only on d.

Proof. For the case $d = 2$, see [36, Prop 1]. Let us then assume that $d \geq 3$ and fix $\theta = \pi/6$. Let m_1 denote the smallest number of cones with apex the origin and aperture 2θ needed to cover the unit ball. (m_1 is finite and depends only on d.) Let W denote the family of unit vectors defining these cones. Define the family

$$\mathcal{U}_{x,\alpha} = \{C_{u,\alpha/2}^\theta(x),\ u \in W\}. \tag{42}$$

The family $\mathcal{U}_{x,\alpha}$ is unavoidable for $\mathcal{E}_{x,\alpha}$, see Figure 7. This follows from the facts that $B(x,\alpha) = \cup_{u\in W} C^{\theta}_{u,\alpha}(x)$ and

$$C^{\theta}_{u,\alpha}(x) \subset \cap_{y\in C^{\theta}_{u,\alpha}(x)} B(y,\alpha). \tag{43}$$

Lemma 2 in [36] provides a proof of (43) in $\mathbb{R}^2$, which can be extended to the general d-dimensional case in a straightforward manner.

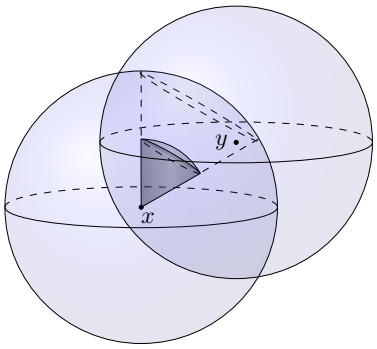

Figure 7: We represent $B(x,\alpha)$ and, in dark gray, the set $C^{\theta}_{u,\alpha/2}(x)$ in $\mathbb{R}^3$. For all $y \in C_{u,\alpha}$, we have that $C^{\theta}_{u,\alpha/2}(x) \subset B(y,\alpha)$. The family $\mathcal{U}_{x,\alpha}$ is unavoidable for $\mathcal{E}_{x,\alpha}$.

Moreover, for each $u \in W$,

$$\mu\left(C^{\theta}_{u,\alpha/2}(x) \cap S\right) = \mu\left(C^{\theta}_{u,\alpha/2}(x)\right), \tag{44}$$

where the equality comes from the fact that $\operatorname{dist}(x,\partial S) > \alpha/2$. And because $B(x,\alpha/2) \subset \bigcup_{u\in W} C^{\theta}_{u,\alpha/2}(x)$, and $\mu(C^{\theta}_{u,\alpha/2}(x))$ is the same for all $u \in W$, we have

$$\mu(B(x,\alpha/2)) \leq \sum_{u\in W} \mu\left(C^{\theta}_{u,\alpha/2}(x)\right) = m_1 \mu\left(C^{\theta}_{u_0,\alpha/2}(x)\right), \tag{45}$$

for any $u_0 \in W$. We conclude with the fact that $\mu(B(x,\alpha/2)) = \omega_d (\alpha/2)^d$, where ω_d denotes the Lebesgue measure of the unit ball in $\mathbb{R}^d$, obtaining the stated bound with $c_1 := \omega_d/(2^d m_1)$. □

Lemma 3. *Under the conditions of Theorem 2, for any $x \in S$ such that $\operatorname{dist}(x, \partial S) \le \alpha/2$, with $0 < \alpha \le r$, there exists a family $\mathcal{U}_{x,\alpha}$ with at most m_2 elements that is unavoidable for $\mathcal{E}_{x,\alpha}$ and satisfies*

$$\mu(U \cap S) \ge c_2 \alpha^{(d-1)/2} \operatorname{dist}(x, \partial S)^{(d+1)/2}, \quad \forall U \in \mathcal{U}_{x,\alpha},$$

where $c_2, m_2 \ge 1$ depends only on d.

Proof. Let $x \in S$ such that $\operatorname{dist}(x, \partial S) \le \alpha/2$. We denote $\rho = \operatorname{dist}(x, \partial S)$. For $d = 2$, see [36, Prop 2]. For $d \ge 3$, let $P_\Gamma x$ be the metric projection[4] of x onto $\Gamma := \partial S$ and η the outward pointing unit normal vector at $P_\Gamma x$. By the fact that S^c satisfies the r-rolling ball condition, we have that $B(P_\Gamma x - r\eta, r) \subset S$. Thus, if $\mathcal{U}_{x,\alpha}$ is an unavoidable family of sets for $\mathcal{E}_{x,\alpha}$, we have that

$$\mu(U \cap S) \ge \mu(U \cap B(P_\Gamma x - r\eta, r)) \tag{46}$$

for all $U \in \mathcal{U}_{x,\alpha}$. Henceforth we assume, without loss of generality, that x is the origin and $\eta = -e_d$, where e_d denotes the d-th canonical basis vector. Then, the problem reduces to defining a suitable family of sets $\mathcal{U}_{0,\alpha}$ unavoidable for $\mathcal{E}_{0,\alpha}$ and giving a lower bound for $\mu(U \cap B((r-\rho)e_d, r))$ independent of $U \in \mathcal{U}_{0,\alpha}$.

We partition $B(0, \alpha)$ into the following two sets, see Figure 8 (left),

$$G_\alpha = \left\{ y \in B(0,\alpha) : \ \langle y, e_d \rangle \ge -\|y\|/2 \right\}, \tag{47}$$

and

$$F_\alpha = \left\{ y \in B(0,\alpha) : \ \langle y, e_d \rangle < -\|y\|/2 \right\}. \tag{48}$$

In order to simplify the notation, we write C_u^θ and $C_{u,h}^\theta$ to refer to $C_u^\theta(x)$ and $C_{u,h}^\theta(x)$ when $x = 0$.

[4] Lemma A.0.6 in [32] states that, if both S and S^c satisfy the r-rolling condition then ∂S has reach $\ge r$ and, therefore, $P_\Gamma x$ is unique whenever $\operatorname{dist}(x, S) < r$.

First, let us consider G_α. Fix $\theta = \pi/6$ and $\gamma \in (0, \pi/6)$, say $\gamma = \pi/7$. There exists a finite family W^G, with m^G unit vectors (depending only on d), with the property that for all $y \in G_\alpha$ there exists $u \in W^G$ such that $C^\theta_{u,\alpha} \subset B(y, \alpha)$ and $\langle u, e_d \rangle \geq -\sin\gamma$, see Figure 8 (right).

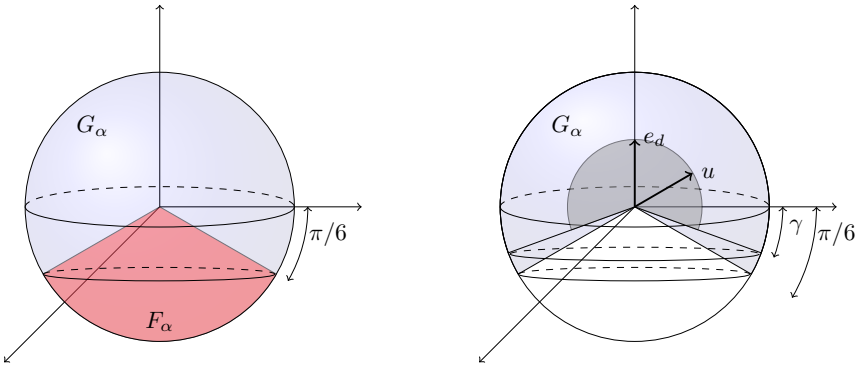

Figure 8: Left: partition of $B(0, \alpha)$ into G_α (in blue) and F_α (in red). Right: For all unitary vector u in the set in dark gray, $\langle u, e_d \rangle \geq -\sin\gamma$.

Let $H_0 = \{x \in \mathbb{R}^d : \langle x, e_d \rangle \geq 0\}$. There is an absolute angle $\tilde{\theta} > 0$ with the property that, for each unit vector u with $\langle u, e_d \rangle \geq -\sin\gamma$ there exists a unit vector $\tilde{u}$ such that $C^{\tilde{\theta}}_{\tilde{u}} \subset C^\theta_u \cap H_0$. Let $\psi = \sqrt{\rho(2r - \rho)}$ and note that

$$H_0 \cap B(0, \psi) \subset B((r - \rho)e_d, r). \tag{49}$$

Hence, for each $u \in W^G$,

$$C^\theta_{u,\alpha} \cap B((r - \rho)e_d, r) \supset C^\theta_{u,\alpha} \cap H_0 \cap B(0, \psi) \supset C^{\tilde{\theta}}_{\tilde{u},\alpha} \cap B(0, \psi) = C^{\tilde{\theta}}_{\tilde{u},\tau}, \tag{50}$$

where $\tau := \min(\psi, \alpha)$.

Using the fact that $\rho \le \alpha \le r$, we have

$$\psi^d = \rho^{\frac{d}{2}}(2r-\rho)^{\frac{d}{2}} = \rho^{\frac{d}{2}}\alpha^{\frac{d}{2}} \ge \rho^{\frac{d+1}{2}}\alpha^{\frac{d-1}{2}} \tag{51}$$

$$\alpha^d = \alpha^{\frac{d+1}{2}}\alpha^{\frac{d-1}{2}} \ge \rho^{\frac{d+1}{2}}\alpha^{\frac{d-1}{2}}, \tag{52}$$

so that

$$\mu\left(B(0,\tau)\right) = \omega_d \tau^d \ge \omega_d\, \alpha^{\frac{d-1}{2}}\rho^{\frac{d+1}{2}}. \tag{53}$$

Also, the ball $B(0,\tau)$ can be covered by a finite number m (depending only on d) of cones $C^{\tilde{\theta}}_{\tilde{u},\tau}$, with varying $\tilde{u}$. Therefore, using the usual argument,

$$\omega_d\, \alpha^{\frac{d-1}{2}}\rho^{\frac{d+1}{2}} \le \mu\left(B(0,\tau)\right) \le m\mu\big(C^{\tilde{\theta}}_{\tilde{u},\tau}\big), \tag{54}$$

and we conclude that

$$\mu\big(C^{\theta}_{u,\alpha} \cap B((r-\rho)e_d, r)\big) \ge L^G \alpha^{\frac{d-1}{2}}\rho^{\frac{d+1}{2}}, \tag{55}$$

where $L^G := \omega_d/m > 0$ only depends on d.

Now, let us consider F_α. First, we define the set

$$C_\ddagger = \left\{x \in \mathbb{R}^d :\ -h_1 \le \langle x, e_d\rangle \le 0\right\} \cap B(-\alpha e_d, \alpha), \tag{56}$$

where $h_1 := \rho(2r-\rho)/(2(r+\alpha-\rho))$; see Figure 9 (left). Note that $C_\ddagger \subset B((r-\rho)e_d, r)$ since, for $x \in C_\ddagger$, $\|x\|^2 \le -2\alpha\langle x, e_d\rangle$ with $\langle x, e_d\rangle \ge -h_1$, which yields $\|x-(r-\rho)e_d\|^2 \le r^2$. Moreover, it can be proved, see [32, Lem 2.4.9], that

$$\mu(C_\ddagger) \ge \frac{\omega_{d-1}}{(d+1)2^{(d-1)/2}}\alpha^{\frac{d-1}{2}}\rho^{\frac{d+1}{2}}. \tag{57}$$

For a unit vector $u \in \mathbb{R}^{d-1}\times\{0\}$, let $Q^\theta_u = \{x = (x_1,\dots,x_d) \in \mathbb{R}^d :\ (x_1,\dots,x_{d-1},0) \in C^\theta_u\}$; see Figure 9 (right).

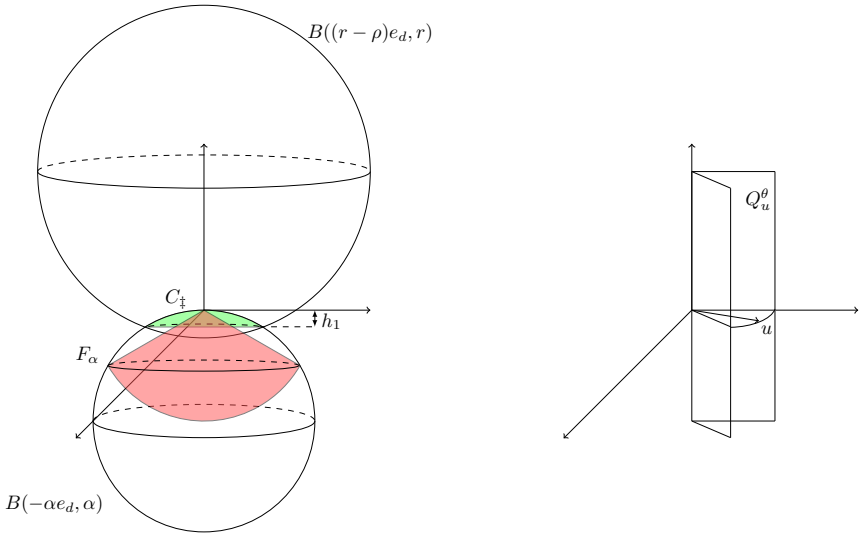

Figure 9: Left, in green set $C_\ddagger$ in $\mathbb{R}^3$. Right, example of set Q_u^θ in $\mathbb{R}^3$.

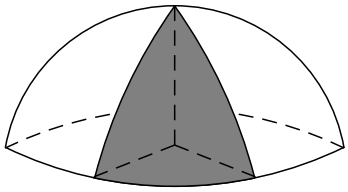

Figure 10: In gray, example of set $Q_u^\theta \cap C_\ddagger$ in $\mathbb{R}^3$.

Next, we define an unavoidable family with sets of the form $Q_u^\theta \cap C_\ddagger$; see Figure 10. We start by noting that there exists a family W^F of m^F unit vectors in $\mathbb{R}^{d-1} \times \{0\}$ (with m^F depending only on d) such that $\mathbb{R}^d = \cup_{u \in W^F} Q_u^\theta$. It can be proved that for all $y \in F_\alpha \cup Q_u^\theta$ there exists $u \in W^F$ such that $Q_u^\theta \cap C_\ddagger \subset B(y, \alpha)$, see [32, Lem 2.4.16]. Then, we use that $C_\ddagger \subset B((r-\rho)e_d, r)$, the fact that $C_\ddagger$ can be covered by the sets $Q_u^\theta \cap C_\ddagger$ with $u \in W^F$, and

(57), to obtain the following sequence of inequalities

$$\mu\left(Q_u^\theta \cap C_\ddagger \cap B((r-\rho)e_d, r)\right) = \mu(Q_u^\theta \cap C_\ddagger) \geq \mu(C_\ddagger)/m^F \geq L^F \alpha^{\frac{d-1}{2}} \rho^{\frac{d+1}{2}}, \tag{58}$$

where $L^F := \omega_{d-1}/(m^F(d+1)2^{(d-1)/2})$ depends only on d.

We finish by defining the family

$$\mathcal{U}_{0,\alpha} = \{C_{u,\alpha}^\theta : u \in W^G\} \cup \{Q_u^\theta \cap C_\ddagger : u \in W^F\}. \tag{59}$$

This completes the proof of the lemma, with $m_2 := m^G + m^F$ and $c_2 = \min(L^G, L^F)$ depending only on d. □

Proof of Theorem 2. Let $S_n = C_\alpha(\mathcal{X}_n)$ and P_X denotes the uniform distribution on S, that is, $P_X(U) = \mu(U \cap S)/\mu(S)$. We have $\mathcal{X}_n \subset S$ (with probability one), which implies that $S_n \subset S$, so that

$$\mathbb{E}_S\left[\mu(S_n \triangle S)\right] = \mathbb{E}_S\left[\mu(S \setminus S_n)\right] = \int_S \mathbb{P}\left(\exists y \in B(x, \alpha_n) : B(y, \alpha) \cap \mathcal{X}_n = \varnothing\right)\mu(\mathrm{d}x), \tag{60}$$

where the second equality is by definition of S_n. In what follows, for each $x \in S$ we choose a finite family $\mathcal{U}_{x,\alpha}$ unavoidable for $\mathcal{E}_{x,\alpha}$. Then, as a consequence of Definition 5, we have

$$\mathbb{P}\left(\exists y \in B(x, \alpha) : B(y, \alpha) \cap \mathcal{X}_n = \varnothing\right) \leq \sum_{U \in \mathcal{U}_{x,\alpha}} (1 - P_X(U))^n \leq \sum_{U \in \mathcal{U}_{x,\alpha}} \exp\left(-nP_X(U)\right). \tag{61}$$

We partition S into two subsets

$$S_1 = \left\{x \in S : \quad \mathrm{dist}(x, \partial S) > \alpha/2\right\}, \tag{62}$$

and

$$S_2 = \left\{x \in S : \quad \mathrm{dist}(x, \partial S) \leq \alpha/2\right\}. \tag{63}$$

For those $x \in S_1$, we choose a family as in Lemma 2, to get

$$\int_{S_1} \sum_{U \in \mathcal{U}_{x,\alpha}} \exp(-nP_X(U))\mu(\mathrm{d}x) \le \int_{S_1} m_1 \exp(-nc_1\alpha^d/\mu(S))\mu(\mathrm{d}x) \tag{64}$$

$$\le \mu(S) m_1 \exp(-nc_1\alpha^d/\mu(S)) \le cn^{-2/(d+1)}, \tag{65}$$

where m_1 and c_1 are defined in Lemma 2, and c only depends on (d, R, α). For those $x \in S_2$, we choose a family as in Lemma 3, and follow the same arguments as in equation (21) in [36], which is based on the change of variables formula for integration, see [6, Thm 16.12], to get

$$\int_{S_2} \sum_{U \in \mathcal{U}_{x,\alpha}} \exp(-nP_X(U))\mu(\mathrm{d}x) \le \int_{S_2} m_2 \exp\left(-nc_2\alpha^{(d-1)/2} \operatorname{dist}(x, \partial S)^{(d+1)/2}/\mu(S)\right) \mu(\mathrm{d}x)$$

$$= \int_0^{\alpha/2} m_2 \exp\left(-nc_2\alpha^{(d-1)/2}\rho^{(d+1)/2}/\mu(S)\right) F'(\rho)\mathrm{d}\rho$$

where m_2 and c_2 are defined in Lemma 3 and $F(\rho) = \mu(x \in S : \operatorname{dist}(x, \partial S) \le \rho)$. By Theorem 5.6 in [17] it is shown that F, as a function of $\rho < r$, is a polynomial in ρ of degree at most d, and the coefficient of degree $d-k$ is proportional to Φ_k, where Φ_k denotes the kth curvature measure associated with ∂S. In Remark 5.10 of Federer [17] it is shown that

$$\sup\left\{|\Phi_k|(T) : T \subset B(0, R),\ \text{reach of } T \ge r\right\} < \infty,$$

where $|\Phi_k|$ is the total variation of Φ_k over T. If S has half-diameter at most R, we can assume without loss of generality that $T = \partial S$ is contained in $B(0, R)$ and, since the reach of T is $\ge r$ by assumption, we have $|F'(\rho)| \le K$ uniformly on S. The constant K only depends on r and R. So, using the change of variable $v = c_2 n\alpha^{(d-1)/2}\rho^{(d+1)/2}/\mu(S)$, we get

$$\int_0^{\alpha/2} m_2 \exp\left(-x_2 n\alpha^{(d-1)/2}\rho^{(d+1)/2}/\mu(S)\right) F'(\rho)d\rho \tag{66}$$

$$\le m_2 K \int_0^{\alpha/2} \exp\left(-c_2 n\alpha^{(d-1)/2}\rho^{(d+1)/2}/\mu(S)\right) d\rho \le c' n^{-2/(d+1)},$$

where c' depends only on (d, r, R, α).

It follows from (65) and (66), that

$$\mathbb{E}_S\left[\mu(C_{\alpha_n}(\mathcal{X}_n) \triangle S)\right] \leq cn^{-2/(d+1)} + c'n^{-2/(d+1)} \leq c''n^{-2/(d+1)}. \tag{67}$$

where c'' depends only on (r, R, d, α). This concludes the proof of (18).

It remains to prove (19). The bound given in (19) can be proved using some results from [43]. In particular, Lemma 3(a) in that paper implies that, for $\epsilon < \alpha/2$, there exists a numerical constant A that depends on (r, R, d) such that

$$\mathbb{P}\left(S \oplus (\alpha - \varepsilon)B_1 \not\subset \mathcal{X}_n \oplus \alpha B_1\right) \leq A\epsilon^{-d}\exp\left(-An\epsilon^{(d+1)/2}\right). \tag{68}$$

Notice that $S \oplus (\alpha - \varepsilon)B_1 \subset \mathcal{X}_n \oplus \alpha B_1$ implies that $(S \oplus (\alpha - \varepsilon)B_1) \ominus \alpha B_1 \subset C_\alpha(\mathcal{X}_n)$. Since S is α-convex, $S = (S \ominus \alpha B_1) \oplus \alpha B_1$ and we get that $(S \oplus (\alpha - \varepsilon)B_1) \ominus \alpha B_1$ can be written as $S \ominus \epsilon B_1$. Therefore, $S \oplus (\alpha - \varepsilon)B_1 \subset \mathcal{X}_n \oplus \alpha B_1$ implies $S \ominus \varepsilon B_1 \subset C_\alpha(\mathcal{X}_n)$. Using (68), we get

$$\mathbb{P}\left(S \ominus \varepsilon B_1 \not\subset C_\alpha(\mathcal{X}_n)\right) \leq A\epsilon^{-d}\exp\left(-An\epsilon^{(d+1)/2}\right). \tag{69}$$

If $S \ominus \varepsilon B_1 \subset C_\alpha(\mathcal{X}_n)$ we have

$$\mu(C_\alpha(\mathcal{X}_n) \triangle S) \leq \mu((S \ominus \varepsilon B_1) \triangle S) \leq \mu(S) - \mu(S \ominus \varepsilon B_1) = O(\varepsilon), \tag{70}$$

uniformly bounded on S in (r, R), see Equation (6) in [43]. Using this bound and (68), we obtain (19). □

References

[1] Arias-Castro, E. and A. Rodríguez-Casal (2017). On estimating the perimeter using the alpha-shape. *Ann. Inst. H. Poincaré Probab. Statist. 53*(3), 1051–1068.

[2] Baddeley, A. and E. Jensen (2004). *Stereology for Statisticians*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. CRC Press.

[3] Baldin, N. (2017). The wrapping hull and a unified framework for estimating the volume of a body. *arXiv preprint arXiv:1703.01658*.

[4] Baldin, N. and M. Reiß (2016). Unbiased estimation of the volume of a convex body. *Stochastic Processes and their Applications 126*(12), 3716 – 3732. In Memoriam: Evarist Giné.

[5] Bárány, I. (2004). Random polytopes in smooth convex bodies: corrigendum. *Mathematika 51*(1-2), 31–31.

[6] Billingsley, P. (1995). *Probability and Measure* (Third ed.). John Wiley and Sons.

[7] Bräker, H. and T. Hsing (1998). On the area and perimeter of a random convex hull in a bounded convex set. *Probab. Theory Related Fields 111*(4), 517–550.

[8] Brown, L. D., T. T. Cai, and A. DasGupta (2001). Interval estimation for a binomial proportion. *Statistical science*, 101–117.

[9] Cholaquidis, A., R. Fraiman, G. Lugosi, and B. Pateiro-López (2016). Set estimation from reflected brownian motion. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 78*(5), 1057–1078.

[10] Clopper, C. J. and E. S. Pearson (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika 26*(4), 404–413.

[11] Cruz-Orive, L. M. and M. García-Fiñana (2005). A review of the article: Comments on the shortcomings of predicting the precision of cavalieri volume estimates based upon assumed measurement functions, by edmund glaser. *Journal of Microscopy 218*(1), 6–8.

[12] Cuevas, A. and R. Fraiman (2010). Set estimation. In *New perspectives in stochastic geometry*, pp. 374–397. Oxford: Oxford Univ. Press.

[13] Cuevas, A., R. Fraiman, and B. Pateiro-López (2012). On statistical properties of sets

fulfilling rolling-type conditions. *Adv. Appl. Probab. 44*(2), 311–329.

[14] Cuevas, A., R. Fraiman, and A. Rodríguez-Casal (2007). A nonparametric approach to the estimation of lengths and surface areas. *Ann. Statist. 35*(3), 1031–1051.

[15] Edelsbrunner, H., D. Kirkpatrick, and R. Seidel (1983a). On the shape of a set of points in the plane. *IEEE Trans. Inf. Theor. 29*(4), 551–559.

[16] Edelsbrunner, H., D. G. Kirkpatrick, and R. Seidel (1983b). On the shape of a set of points in the plane. *IEEE Trans. Inform. Theory 29*(4), 551–559.

[17] Federer, H. (1959). Curvature measures. *Trans. Amer. Math. Soc. 93*, 418–491.

[18] Friedman, J., T. Hastie, and R. Tibshirani (2001). *The elements of statistical learning*, Volume 1. Springer series in statistics Springer, Berlin.

[19] Gayraud, G. (1997). Estimation of functionals of density support. *Mathematical Methods of Statistics 6*(1), 26–46.

[20] Getz, W. M., S. Fortmann-Roe, P. C. Cross, A. J. Lyons, S. J. Ryan, and C. C. Wilmers (2007). Locoh: Nonparameteric kernel methods for constructing home ranges and utilization distributions. *PLOS ONE 2*(2), 1–11.

[21] Gilja, O. H., T. Hausken, A. Berstad, and S. Ødegaard (1999). Measurements of organ volume by ultrasonography. *Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine 213*(3), 247–259.

[22] Hall, P. and J. Ziegel (2011). Distribution estimators and confidence intervals for stereological volumes. *Biometrika 98*(2), 417–431.

[23] Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc. 58*, 13–30.

[24] Jiménez, R. and J. E. Yukich (2011). Nonparametric estimation of surface integrals. *Ann. Statist. 39*(1), 232–260.

[25] Kim, J.-C. and A. Korostelev (2000). Estimation of smooth functionals in image

models. *Mathematical Methods of Statistics 9*(2), 140–159.

[26] Korhonen, L., J. Vauhkonen, A. Virolainen, A. Hovi, and I. Korpela (2013). Estimation of tree crown volume from airborne lidar data using computational geometry. *International Journal of Remote Sensing 34*(20), 7236–7248.

[27] Korostelëv, A. P. and A. B. Tsybakov (1993). *Minimax theory of image reconstruction*, Volume 82 of *Lecture Notes in Statistics*. New York: Springer-Verlag.

[28] Liang, J., H. Edelsbrunner, P. Fu, P. V. Sudhakar, and S. Subramaniam (1998). Analytical shape computation of macromolecules: I. molecular area and volume through alpha shape. *Proteins: Structure, Function, and Bioinformatics 33*(1), 1–17.

[29] Mammen, E. and A. B. Tsybakov (1995). Asymptotical minimax recovery of sets with smooth boundaries. *Ann. Statist. 23*(2), 502–524.

[30] Opsomer, J. D., F. J. Breidt, G. G. Moisen, and G. Kauermann (2007). Model-assisted estimation of forest resources with generalized additive models. *Journal of the American Statistical Association 102*(478), 400–409.

[31] Ozenne, B., F. Subtil, L. Østergaard, and D. Maucort-Boulch (2015). Spatially regularized mixture model for lesion segmentation with application to stroke patients. *Biostatistics 16*(3), 580–595.

[32] Pateiro-Lopez, B. (2008). *Set estimation under convexity type restrictions*. Ph. D. thesis, Universidad de Santiago de Compostela.

[33] Pateiro-López, B. and A. Rodríguez-Casal (2008). Length and surface area estimation under smoothness restrictions. *Adv. in Appl. Probab. 40*(2), 348–358.

[34] Pateiro-López, B. and A. Rodríguez-Casal (2009). Surface area estimation under convexity type assumptions. *J. Nonparametr. Stat. 21*(6), 729–741.

[35] Pateiro-López, B. and A. Rodrıguez-Casal (2010). Generalizing the convex hull of a sample: The r package alphahull. *Journal of Statistical software 34*(5), 1–28.

[36] Pateiro-López, B. and A. Rodríguez-Casal (2013). Recovering the shape of a point cloud in the plane. *TEST 22*(1), 19–45.

[37] Perkal, J. (1956). Sur les ensembles ε-convexes. *Colloq. Math. 4*, 1–10.

[38] Powell, R. (2000). Animal home ranges and territories and home range estimators. In *Research techniques in animal ecology: controversies and consequences*, pp. 64–110.

[39] Rényi, A. and R. Sulanke (1964). Über die konvexe Hülle von n zufällig gewählten Punkten. II. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete 3*, 138–147 (1964).

[40] Rodríguez-Casal, A. (2007). Set estimation under convexity type assumptions. *Ann. Henri Poincaré 43*(6), 763–774.

[41] Rodríguez-Casal, A. and P. Saavedra-Nieves (2016). A fully data-driven method for estimating the shape of a point cloud. *ESAIM: Probability and Statistics 20*, 332–348.

[42] Tsybakov, A. B. (2009). *Introduction to nonparametric estimation*. Springer Series in Statistics. New York: Springer. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats.

[43] Walther, G. (1997). Granulometric smoothing. *The Annals of Statistics 25*(6), pp. 2273–2299.

[44] Walther, G. (1999). On a generalization of Blaschke's rolling theorem and the smoothing of surfaces. *Math. Methods Appl. Sci. 22*(4), 301–316.

[45] Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association 22*(158), 209–212.

[46] Yu, B. (1997). Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam*, pp. 423–435. New York: Springer.